Visual Mismatch and Predictive Coding: A Computational Single-Trial ERP Study

Stefanics, Gabor ; Heinzle, Jakob ; Horváth, András Attila ; Stephan, Klaas Enno

Abstract: Predictive coding (PC) posits that the brain uses a generative model to infer the environmental causes of its sensory data and uses precision-weighted prediction errors (pwPEs) to continuously update this model. While supported by much circumstantial evidence, experimental tests grounded in formal trial-by-trial predictions are rare. One partial exception is event-related potential (ERP) studies of the auditory mismatch negativity (MMN), where computational models have found signatures of pwPEs and related model-updating processes. Here, we tested this hypothesis in the visual domain, examining possible links between visual mismatch responses and pwPEs. We used a novel visual "roving standard" paradigm to elicit mismatch responses in humans (of both sexes) by unexpected changes in either color or emotional expression of faces. Using a hierarchical Bayesian model, we simulated pwPE trajectories of a Bayes-optimal observer and used these to conduct a comprehensive trial-by-trial analysis across the time × sensor space. We found significant modulation of brain activity by both color and emotion pwPEs. The scalp distribution and timing of these single-trial pwPE responses were in agreement with visual mismatch responses obtained by traditional averaging and subtraction (deviant-minus-standard) approaches. Finally, we compared the Bayesian model to a more classical change model of MMN. Model comparison revealed that trial-wise pwPEs explained the observed mismatch responses better than categorical change detection. Our results suggest that visual mismatch responses reflect trial-wise pwPEs, as postulated by PC. These findings go beyond classical ERP analyses of visual mismatch and illustrate the utility of computational analyses for studying automatic perceptual processes.

DOI: https://doi.org/10.1523/JNEUROSCI.3365-17.2018

JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Visual mismatch and predictive coding: A computational single-trial ERP study

Gabor Stefanics[1,2], Jakob Heinzle[1], András Attila Horváth[3] and Klaas Enno Stephan[1,4,5]

[1]*Translational Neuromodeling Unit (TNU), University of Zurich & ETH Zurich, 8032 Zurich, Switzerland,*
[2]*Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, 8006 Zurich, Switzerland*
[3]*National Institute of Clinical Neurosciences, Department of Neurology, National Brain Research Program, 1145 Budapest, Hungary*
[4]*Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, WC1N 3BG London, UK.*
[5]*Max Planck Institute for Metabolism Research, 50931 Cologne, Germany.*

DOI: 10.1523/JNEUROSCI.3365-17.2018

Received: 28 November 2017

Revised: 12 February 2018

Accepted: 13 March 2018

Published: 26 March 2018

Author contributions: G.S. wrote the first draft of the paper; G.S. and K.E.S. designed research; G.S. and A.A.H. performed research; G.S. contributed unpublished reagents/analytic tools; G.S. analyzed data; G.S., J.H., and K.E.S. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

We acknowledge support by the University of Zurich (KES), the René and Susanne Braginsky Foundation (KES), and the Clinical Research Priority Program “Multiple Sclerosis” (GS, KES).

Corresponding author: Gabor Stefanics, Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Wilfriedstrasse 6, CH - 8032 Zürich, e-mail: stefanics@biomed.ee.ethz.ch

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.3365-17.2018

Title:

Visual mismatch and predictive coding: A computational single-trial ERP study

Abbreviated title:

Visual mismatch and predictive coding

Gabor Stefanics[1,2], Jakob Heinzle[1], András Attila Horváth[3], Klaas Enno Stephan[1,4,5]

[1]Translational Neuromodeling Unit (TNU), University of Zurich & ETH Zurich, 8032 Zurich, Switzerland,

[2]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, 8006 Zurich, Switzerland

[3]National Institute of Clinical Neurosciences, Department of Neurology, National Brain Research Program, 1145 Budapest, Hungary

[4]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, WC1N 3BG London, UK.

[5] Max Planck Institute for Metabolism Research, 50931 Cologne, Germany.

Corresponding author:

Gabor Stefanics, Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Wilfriedstrasse 6, CH - 8032 Zürich, e-mail: stefanics@biomed.ee.ethz.ch

Number of pages: 23

Number of figures: 6, tables: 1, multimedia: 0 and 3D models: 0

Number of words for Abstract: 250/250, Introduction: 633/650, and Discussion: 1500/1500

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements:

We acknowledge support by the University of Zurich (KES), the René and Susanne Braginsky Foundation (KES), and the Clinical Research Priority Program “Multiple Sclerosis” (GS, KES).

Abstract

Predictive coding (PC) posits that the brain employs a generative model to infer the environmental causes of its sensory data and uses precision-weighted prediction errors (pwPE) to continuously update this model. While supported by much circumstantial evidence, experimental tests grounded in formal trial-by-trial predictions are rare. One partial exception are event-related potential (ERP) studies of the auditory mismatch negativity (MMN), where computational models have found signatures of pwPEs and related model-updating processes.

Here, we tested this hypothesis in the visual domain, examining possible links between visual mismatch responses and pwPEs. We used a novel visual 'roving standard' paradigm to elicit mismatch responses in humans (of both sexes) by unexpected changes in either color or emotional expression of faces. Using a hierarchical Bayesian model, we simulated pwPE trajectories of a Bayes-optimal observer and used these to conduct a comprehensive trial-by-trial analysis across the time×sensor space. We found significant modulation of brain activity by both color and emotion pwPEs. The scalp distribution and timing of these single-trial pwPE responses were in agreement with visual mismatch responses obtained by traditional averaging and subtraction (deviant-minus-standard) approaches. Finally, we compared the Bayesian model to a more classical change detection (CD) model of MMN. Model comparison revealed that trial-wise pwPEs explained the observed mismatch responses better than categorical change detection.

Our results suggest that visual mismatch responses reflect trial-wise pwPEs, as postulated by PC. These findings go beyond classical ERP analyses of visual mismatch and illustrate the utility of computational analyses for studying automatic perceptual processes.

Significance Statement (120/120)

Human perception is thought to rely on a predictive model of the environment which is updated via precision-weighted prediction errors (pwPE) when events violate expectations. This "predictive coding" view is supported by studies of the auditory mismatch negativity brain potential. However, it is less well known whether visual perception of mismatch relies on similar processes. Here we combined computational modeling and electroencephalography to test whether visual mismatch responses reflected trial-by-trial pwPEs. Applying a Bayesian model to series of face stimuli that violated expectations about color or emotional expression, we found significant modulation of brain activity by both color and emotion pwPEs. A categorical change detection model performed less convincingly. Our findings support the predictive coding interpretation of visual mismatch responses.

Keywords:

61 Bayesian inference, computational modeling, EEG, visual MMN (vMMN), precision-weighted prediction
62 error (pwPE)

63

Introduction

According to predictive coding (PC), sensory systems operate under hierarchical Bayesian principles in order to infer the causes of their sensory inputs. This rests on message passing among hierarchically related neuronal populations: each level sends predictions to the level below and receives precision-weighted prediction errors (pwPEs) in return which serve to update predictions (Rao and Ballard, 1999; Friston, 2005; Hohwy, 2013; Clark, 2015). This process of perceptual inference is optimized by learning, where pwPEs to repeated sensory events are explained away with increasing efficiency, mediated by plastic changes in synaptic connections of the sensory circuits (Friston, 2005; Baldeweg, 2006).

Perceptual learning experiments often use stimulus repetition to establish expectations. An experimental protocol frequently used to study implicit perceptual learning in audition is the 'roving standard' paradigm (Haenschel et al., 2005; Garrido et al., 2008; Costa-Faidella et al., 2011a,b; Schmidt et al., 2013; Moran et al., 2013; Auksztulewicz and Friston, 2015; Komatsu et al., 2015; Takaura and Fujii, 2016). This repeats a stimulus several times before unpredictably switching to a different stimulus train. This paradigm is frequently used to elicit the "mismatch negativity" (MMN), an event-related potential (ERP) that signals violations of statistical regularities during perceptual learning. Although the MMN was primarily investigated in the auditory modality (for reviews, see Näätänen et al., 2010, 2012) there is increasing evidence for MMN also in the visual modality (for reviews, see Stefanics et al., 2014; Kremláček et al., 2016).

Since its discovery, the MMN response has been interpreted in different ways. First, the "memory-trace" or "change-detection" hypothesis (Näätänen et al., 1989, 1993; Schröger, 1998) conceptualized the MMN as a brain response signaling the difference between the immediate history of the stimulus sequence and a novel stimulus. Later, this interpretation was followed by the "regularity violation" hypothesis (Winkler, 2007), according to which the MMN signals a difference between the current stimulus and expectations based on prior information which might not only represent a sensory memory trace but also more complex or abstract rules extracted from regular relationships between preceding stimuli, e.g., conditional probabilities (e.g., Paavilainen et al., 2007; Stefanics et al., 2009, 2011); for a review see Paavilainen, 2013). This interpretation is compatible with the most recent view of the MMN as an expression of pwPEs during PC (Friston, 2005; Baldeweg, 2006; Stephan et al., 2006; Wacongne et al., 2011; Lieder et al., 2013a; Stefanics et al., 2015). In fact, a PC view of MMN can be seen as mathematically formalizing ideas already inherent to the earlier "regularity violation" hypothesis.

The PC interpretation of MMN is supported by much, albeit mostly indirect, experimental evidence (e.g., Garrido et al., 2007, 2013, 2017; Stefanics and Czigler, 2012; Phillips et al., 2015; Auksztulewicz and Friston, 2016; Chennu et al., 2016). By contrast, experimental studies based on formal trial-by-trial computational quantities are rare, almost entirely restricted to the auditory domain, and typically focused on specific sensors or time windows (Lieder et al., 2013b; Kolossa et al., 2015; Jepma et al.,

2016). Here, we go beyond previous investigations and use a Bayesian model (the Hierarchical Gaussian Filter, HGF) to examine whether visual mismatch responses reflect pwPEs, a hallmark of PC.

Specifically, our paradigm used a "roving" design in which two features of human faces were altered probabilistically and orthogonally: color and emotional expression. We used the HGF to generate pwPE trajectories and tested the implication by PC, that trial-by-trial brain activity would reflect these computational quantities. In addition, we applied a trial-wise change detection (CD) model (cf. Lieder et al., 2013b) and evaluated the explanatory power of both hypotheses by statistical model comparison. Finally, we analyzed visual mismatch responses (aka visual mismatch negativity (vMMN) responses; for reviews, see Stefanics et al., 2014; Kremláček et al., 2014) obtained with traditional averaging and subtraction methods, and compared the results to those obtained by modeling.

Methods

Ethics Statement

The experimental protocol was approved by the Cantonal Ethics Commission of Zurich (KEK 2011-0239/3). Written informed consent was obtained from all participants after the procedures and risks were explained. The experiments were conducted in compliance with the Declaration of Helsinki.

Subjects

Thirty-nine neurologically normal subjects volunteered in this experiment. One subject's data was excluded due to excessive blinks, and four subjects' data were rejected because of bridges between electrodes due to conductive gel. The final sample comprised 34 subjects (mean age=23.88ys, SD=3.56ys, 17 females, 33 right-handed). All subjects had normal or corrected-to-normal vision.

Paradigm

We used a multi-feature visual 'roving standard' paradigm to elicit mismatch responses (PEs) by rare changes either in color (red, green), or emotional expression (happy, fearful) of human faces, or both. Roving paradigms have often been used to elicit automatic sensory expectations in the auditory modality by manipulating stimulus probabilities (Haenschel et al., 2005; Garrido et al., 2008; Moran et al., 2013; Auksztulewicz and Friston, 2015). Here, we presented four types of visual stimuli (green fearful, green happy, red fearful, and red happy faces). Hence, each stimulus type could violate expectations either about the color or emotional expression of faces (or both). Importantly, this allowed us to study brain responses to stimuli that were physically identical but differed in whether color or emotion regularities were violated. Faces were presented in four peripheral quadrants of the screen (Fig. 1A). Each stimulus type was presented with an equal overall probability (p=0.25) during the experiment. After 5-9 presentations each stimulus type was followed by any of the other three types with equal overall transition probabilities (Fig. 1B). Participants engaged in a central detection task that required speeded button-presses to changes of the fixation cross. Reaction times were recorded. The experiment consisted of 14 blocks, each lasting about 8 minutes. A short training session preceded the EEG recording.

Face stimuli, ten female and ten male Caucasian models, were selected from the Radboud Faces Database (Langner et al., 2010; www.rafd.nl) based on their high percentage of agreement on emotion categorization (98% for happy, 92% for fearful faces). To control low-level image properties, we used the SHINE toolbox (Willenbockel et al., 2010) to equate luminance and spatial frequency content of grayscale images of the selected happy and fearful faces. The resulting images were used to create the colored stimuli.

144

145 ---------- Figure 1 around here ----------

146

147

148 Faces were presented on a CRT monitor on a dark-grey background at a viewing distance of 1m. The
149 width and height of each face subtended 3.8° and 5.4° visual angle, respectively. The horizontal and
150 vertical distance of the center of the face stimuli from the center of the screen was 3.15°. To avoid
151 potential local adaptation effects, each stimulus panel consisted of four faces with different identity
152 (two females, two males) and the presentation order of the faces with different identity was
153 randomized with the restriction that a face with the same identity was not presented in adjacent trials.
154 Each face was presented with the same probability over the experiment. Stimuli were presented for 200
155 ms, followed by a random inter-stimulus interval of 600-700 ms during which only the fixation cross was
156 present. Stimuli were presented using Cogent2000 (http://www.vislab.ucl.ac.uk/Cogent/index.html).

157 EEG recording and preprocessing

158 During the experiment, participants sat in a comfortable chair in an electromagnetically shielded, sound-
159 attenuated, dimly lit room. Continuous EEG was recorded from 0.016 Hz with a low-pass filter at 100 Hz
160 using a QuickAmp amplifier (BrainProducts, Gilching, Germany). The high-density 128-channel electrode
161 caps had an equidistant hexagonal layout and covered the whole head. EEG was referenced against the
162 common average potential; the ground electrode was placed on the right cheek. Electrodes above the
163 eyes and near the left and right external canthi were used to monitor eye movements. Data were
164 digitized at 24 bit resolution and a sampling rate of 500 Hz and filtered off-line between 0.5 and 30 Hz
165 using zero-phase shift infinite-impulse response (IIR) Butterworth filter. Built-in and self-developed
166 functions as well as the freeware SPM12 toolbox (v6470, RRID: SCR_007037; Litvak et al., 2011) in the
167 Matlab development environment (MathWorks, Natick, MA) were used for subsequent off-line data
168 analyses. Electrode positions and fiducials were digitized for each subject using an infrared light-based
169 measurement system and Xensor software (ANT B.V., Enschede, The Netherlands).

170 Epochs extending -100 ms before to 500 ms after stimulus onset were extracted from the continuous
171 EEG. Epochs were baseline corrected using the 100 ms pre-stimulus period. A topography-based artifact
172 correction method (Berg and Scherg, 1994) implemented in SPM12 was used to correct for eye-blink
173 and eye-movement artifacts. Electrode positions were used to co-register EEG data to a canonical MRI
174 template to calculate a forward model to define topographies of blink and eye-movement artifacts
175 which were removed from the epoched data. To avoid other potential artifacts, epochs with values
176 exceeding ±100 μV on any EEG channel were rejected from the analysis.

177 Modeling belief trajectories

We used the Hierarchical Gaussian Filter (Mathys et al., 2011; Mathys et al., 2014) to simulate computational trajectories in order to create parametric regressors for the general linear model (GLM) analysis. The HGF is a generative (Bayesian) model of perceptual inference and learning that represents a variant of PC in the temporal domain and that has been used in several recent studies to investigate hierarchical PE responses in the brain (Iglesias et al., 2013; Hauser et al., 2014; Schwartenbeck et al., 2015; Vossel et al., 2015; Lawson et al., 2017; Powers et al., 2017). It is implemented in the freely available open source software TAPAS (http://www.translationalneuromodeling.org/tapas). The HGF consists of a perceptual and a response model, representing a Bayesian observer who receives a sequence of inputs (stimuli) and generates behavioral responses. The perceptual model describes a hierarchical belief updating process, i.e., inference about hierarchically related environmental states that give rise to sensory inputs. In our MMN paradigm the ERP-eliciting face stimuli did not require a behavioral response. Therefore, we used only the perceptual model to simulate belief trajectories about external states, e.g., the occurrence of a red vs. green, or a fearful vs. happy face, without specifying a decision model.

---------- Figure 2 around here ----------

The HGF (Fig. 2A) describes how hidden states (x) of the world generate sensory inputs (u). Model inversion infers these hidden states from sensory inputs; this is equivalent to updating the beliefs across the HGF hierarchy. Here, we used a two-level version of the HGF (based on toolbox v2.2) where we eliminated the third level from the most commonly used hierarchy. This model assumes a stable volatility over the time-course of the experiment, which is in line with the stimulus sequence. The first level of the model represents a sequence of beliefs about stimulus occurrence x_1. This corresponds to beliefs about environmental states, i.e., whether a green vs. red face, or a happy vs. fearful face was presented. The second level represents the current belief of the probability that a given stimulus occurs, i.e., the tendency x_2 towards a given feature (e.g., the conditional probability of seeing a red face vs. a green face, given the previous stimulus).

The model assumes that environmental hidden states evolve as a Gaussian random walk, such that their variance depends on the state at the next higher level (Mathys et al., 2011, 2014):

$$p(x_1|x_2) = s(x)^{x_1}(1 - s(x_2))^{1-x_1} = \text{Bernoulli}(x_1; s(x_2)) \quad (1)$$

$$p\left(x_2^{(k)}\middle|x_2^{(k-1)}, x_3^{(k)}\right) = N(x_2^{(k)};\ x_2^{(k-1)}, \exp(\omega)) \quad (2)$$

where k is a trial index and s is a sigmoid function

$$s(x) = \frac{1}{1+\exp(-x)} \quad (3)$$

At the second level, the top-level in our implementation (equation 2), the step size between consecutive time steps depends on ω.

Exact Bayesian inversion requires analytically intractable integrations, therefore the HGF relies on a quadratic approximation to the variational energies. The variational inversion of the model provides a set of analytical update equations, which update trial-by-trial the model's estimates of the state variables. Importantly, every belief within the model is updated after each trial, leading to trial-by-trial trajectories of these hidden quantities. The update rules share a general form across the model's hierarchy: at any level i the update of the posterior mean $\mu_i^{(k)}$ of the state x_i that represents the belief on trial k is proportional to the precision-weighted PE $\varepsilon_i^{(k)}$. This weighted PE is the product of the PE $\delta_{i-1}^{(k)}$ from the level below and a precision ratio $\psi_i^{(k)}$:

$$\mu_i^{(k-1)} - \mu_i^{(k)} \propto \psi_i^{(k)} \delta_{i-1}^{(k)} = \varepsilon_i^{(k)} \quad (4)$$

The update equations of the hidden states of the HGF (level 2 here) have a general structure similar to those of classical reinforcement or associative learning models, such as the Rescorla-Wagner learning (Rescorla and Wagner, 1972):

$$\text{prediction}^{(k)} = \text{prediction}^{(k\text{-}1)} + \text{learning rate} \times \text{prediction error} \quad (5)$$

We focus our EEG analysis on the pwPE on the second level ε_2, which drives learning about the probability of the stimulus. Here, we provide a brief description of the nature of this quantity. For a detailed and more general derivation of mathematical details see Mathys et al. (2011). The update equation of the mean of the second level is:

$$\mu_2^{(k)} = \mu_2^{(k-1)} + \sigma_2^{(k)}(\mu_1^{(k)} - s(\mu_2^{(k-1)})) \quad (6)$$

where the last term is the PE $(\mu_1^{(k)} - s(\mu_2^{(k-1)}))$ at the first level weighted by the precision term $\sigma_2^{(k)}$. This pwPE updates beliefs at the second level. The precision weight is also updated with every trial and can be regarded as equivalent to a dynamic learning rate in reward learning models (cf. Preuschoff and Bossaerts, 2007). Thus, $\varepsilon_2^{(k)}$ is not simply a scaled version of $\delta_1^{(k)}$.

We computed trajectories of pwPEs (with separate models for color and emotion stimuli) assuming a Bayes-optimal observer. For this, we modeled belief trajectories by estimating the parameters that would lead to minimal surprise about the stimuli. We determined these Bayes-optimal perceptual parameters by inverting the perceptual model based on the stimulus sequence alone and under a predefined prior (the standard in the HGF toolbox). Thus, our modeled observer was the same for all participants and was optimal under its prior beliefs encoded by the parameters that controlled the evolution of the estimated hidden states (Mathys et al., 2011). These trajectories capture the evolution of pwPEs – a hallmark of predictive coding – over each and every trial, peaking when a stimulus

represented a change relative to previous stimuli, and subsiding over following repetitions (Fig. 2B). These model-derived trajectories can thus be used as quantitative regressors in a GLM single-trial analysis of EEG data, without the need to manually label trials as "deviants" or "surprising". We used the absolute value of pwPE traces for the four stimulus types (Fig. 2B) to create regressors that entered the GLM which we estimated for each participant.

Space × time SPM analysis and model comparison

Single-trial sensor data were downsampled to 250 Hz and converted to scalp × time images for statistical analysis. Data were interpolated to create a 32×32 pixel scalp map for each time-point in the poststimulus 50-500 ms interval. The time dimension consisted of 113 samples (of 4 ms) in each trial. Images were stacked to create a 3D space-time image volume which was smoothed with a Gaussian kernel (full-width at half-maximum (FWHM)=[16mm 16mm 16ms]) in accordance with the assumptions of Random Field Theory (Worsley et al., 1996; Kiebel and Friston, 2004).

We performed statistical parametric mapping across the time×sensor space, using two separate GLMs incorporating regressors from the HGF and from a more classical change detection model (CD; see Lieder et al. 2013), respectively. Both models make trial-by-trial predictions about mismatch responses, but differ in the exact form of the ensuing trajectories (HGF: gradually changing pwPEs; CD: categorical changes). For the HGF-based GLM, we included the four stimulus types as main regressors, and color-pwPE and emotion-pwPE as parametric modulators for each stimulus type. For the GLM based on the CD model, we included the four stimulus types as main regressors, and stick functions as parametric modulators for each stimulus type on those trials when a change occurred in the stimulus sequence. The GLMs were estimated for each participant individually.

Group level analyses used F-tests to find scalp time-points where single-trial ERPs were significantly modulated by pwPEs. The resulting statistical parametric maps (SPM) were family-wise error (FWE) corrected for multiple comparisons at the cluster level ($p<0.05$; with a cluster defining threshold of $p<0.001$, as recommended by (Flandin and Friston, 2016) using Random Field Theory. Similar preprocessing and statistical procedures have been applied elsewhere (e.g., Henson et al., 2008; Garrido et al., 2013; Auksztulewicz and Friston, 2015).

In order to compare the two models formally, we used the Bayesian Information Criterion (BIC) (Schwarz, 1978) approximation to the log model evidence (LME), separately for each participant. Under Gaussian noise (as assumed by the GLM), this leads to an approximation that is a function of the residual sum of squares (RSS):

$$LME \simeq \frac{1}{2} n \ln\left(\frac{RSS}{n}\right) + \frac{1}{2} k \ln(n) \quad (7)$$

where *n* is the number of data points and *k* is the number of parameters estimated by the model. Notably, in our case, *n* and *k* are the same in both models. Hence, the difference between the LMEs and, therefore, model comparison depends only on the logarithm of the RSS, i.e. model fit.

In order to perform model comparison at the group level, we computed the logarithm of the group Bayes factor (GBF; Stephan et al., 2007) for each voxel, i.e., the sum of ΔLME (between models) across subjects. This corresponds to a fixed effects group-level Bayesian model selection (BMS; Stephan et al. 2009) procedure and was done both within a functionally defined mask (of voxels showing mismatch responses under both models) as well as on all voxels in the 3D space-time image volume (to perform an unrestricted comparison). The mask comprised all voxels from the SPM analyses where, either for color or emotion changes, both the pwPE and the CD model ("logical AND" conjunction) had yielded a significant whole-brain corrected effect. We then used a non-parametric Wilcoxon signed rank test to assess the null hypothesis of zero median for ΔLME across all voxels.

Traditional ERP analysis

In addition to the model-based approach, we studied mismatch effects using traditional analysis methods by comparing ERP responses to deviants and standards. Deviants were defined as the first stimulus representing a change either in color or in emotion in the stimulus sequence relative to the preceding stimulus; standards were defined as responses to the same stimulus after five repetitions (the 6^{th} presentation of the same stimulus in a row; e.g., Garrido et al., 2008). Thus we compared responses to physically identical stimuli.

Deviant and standard ERP amplitudes were tested for significant MMN response at three posterior region of interest (ROI) at the left occipito-temporal, middle occipital, and right occipito-temporal regions. Regions and time windows for analysis were selected based on prior literature for color (Czigler et al., 2002; Kimura et al., 2006; Thierry et al., 2009; Czigler and Sulykos, 2010; Müller et al., 2010; Mo et al., 2011; Stefanics et al., 2011) and emotion (Zhao and Li, 2006; Astikainen and Hietanen, 2009; Kimura et al., 2012; Stefanics et al., 2012; Astikainen et al., 2013; Csukly et al., 2013; Kreegipuu et al., 2013) changes. Prior studies measured ERP amplitudes consistently at posterior occipital, temporal, and parietal regions. However, the time windows selected for analysis varied remarkably across studies in the 100-500 ms range, therefore we adopted a flexible approach and measured ERP amplitudes to deviants and standards in twelve 32 ms long consecutive intervals in the 100-484 ms range. The effect of stimulus type on evoked responses was tested by a three-way analysis of variance (ANOVA) of Stimulus type (Deviant vs. Standard) × ROI (Left vs. Middle vs. Right) × Interval (12 intervals). Greenhouse–Geisser correction of the degrees of freedom was applied where appropriate, ε values are provided in the results. Significant main effects and interactions were further specified by Tukey HSD (Honestly Significant Difference) post-hoc tests.

Results

Trial-by-trial pwPE results (Bayesian model)

Our analysis across the time×sensor space demonstrated strong correlations between model-based pwPE trajectories, ε2, and the single-trial ERPs (Fig. 3A), both for color and emotion. Details of test statistics are given in Table 1. F-tests revealed significant activations for color pwPEs in several space × time clusters (scalp areas and time intervals). The earliest significant interval was found between 180-255 ms at left and right posterior regions (Fig. 3B), corresponding to a negative potential (Fig. 5B), as well as a fronto-central positivity in a corresponding time window. We observed further correlations at a middle occipital area in the 320-430 ms interval corresponding to a positive potential, as well as negativity in a similar time window with fronto-central dominance. Furthermore, we found a middle occipito-parietal interval in the 430-500 ms time window corresponding to a positive potential, with corresponding fronto-central negativity in a similar time window.

---------- Table 1 around here ----------

For emotion pwPEs, F-tests revealed significant activations in two space × time clusters (Fig. 3C). The earliest effects for emotion PEs were observed at a right occipito-temporal area in the 170-214 ms interval, followed by positivity at the left occipito-temporal scalp region in the 405-455 ms interval (Fig. 3D).

---------- Figure 3 around here ----------

To demonstrate the relationship between the model-based pwPE parameter estimates for color changes and the MMN obtained from ERP data using traditional averaging and subtraction methods, we plotted all raw single-trials sorted in an increasing order according to the trial-wise parameter estimates (Fig. 4A, B). The relationship between the computational quantities of pwPE estimates and raw data is apparent in plots showing the trial-wise ERP amplitudes (Fig. 4C) in the time windows where statistical parametric mapping yielded significant results. Calculating the mean ERP for the 10% of trials with the lowest and highest pwPE estimates, respectively, reveals characteristic ERP waveforms (Fig. 4D) that clearly differ in time intervals where classical deviant-minus-standard differences (early MMN, and late positivity) have been reported previously. A similar, although less robust relationship between model-based pwPE parameter estimates for emotion changes and the ERP data is shown in Fig. 4E-H.

---------- Figure 4 around here ----------

Comparison to the change detection (CD) model

In order to assess, whether the pwPE traces provided any advantage in modeling the EEG data compared to a classical CD model, we performed statistical model comparison. This was based on computing voxel-wise log group Bayes factors (using a BIC-approximation to the group-level log model evidence difference ΔLME), as described in the Methods section. Figure 5 shows that the large majority of the voxels within a functionally defined mask showed strong evidence for the pwPE model (median ΔLME=29.14, mean ΔLME= 33.48, sd=37). ΔLMEs within the whole 3D space-time volume showed very similar results (median ΔLME=29.31, mean ΔLME=31.34, sd=34.86). Notably, a difference in LME >5 is considered as very strong evidence in favor of the superior model (Kass and Raftery, 1995).

To characterize the distribution of ΔLME values more formally, we performed null hypothesis testing. An initial one-sample Kolmogorov-Smirnov test indicated that the distributions of ΔLME for voxels within our functionally defined mask ($D=0.78$, $p<10^{-5}$) as well as for the whole 3D space-time volume ($D=0.79$, $p<10^{-5}$) was not Gaussian. A non-parametric Wilcoxon signed rank test was used to test the null hypothesis of zero median for the ΔLME. The results showed that the median ΔLME was significantly different from zero ($Z=-70.63$, $p<10^{-5}$) for voxels within the mask, as well as for voxels within the whole volume ($Z=-213.10$, $p<10^{-5}$). Distributions of ΔLME values within the significance mask and the entire 3D space-time volume are shown in Figure 5. These results indicate the superiority of the Bayesian model over the CD model and suggest that visual mismatch responses are better explained by pwPEs than by categorical change indices.

---------- Figure 5 around here ----------

Traditional ERP results

Figures 6A and 6B show grand-average ERPs to color deviant and standard as well as to emotion deviant and standard stimuli, respectively, at occipito-temporal/occipital ROIs. Stimuli evoked the canonical P1, N1/N170 and P2 components. Deviant-minus-standard difference waves show a typical visual mismatch negativity around 200 ms for color changes, followed by a positive potential after 300 ms. ERP

waveforms obtained with traditional averaging and subtraction methods reveal a smaller negativity for emotion changes peaking before 200 ms in the right ROI followed by a positivity after 400 ms that is most robust on the left ROI (Fig. 6C, D).

The ANOVA of the amplitude values for color deviants and standards yielded a significant interaction of Stimulus type × Interval ($F(11,363)=14.491$, $p<0.00001$, $\varepsilon=0.369$, $\eta^2=0.305$). A post-hoc Tukey test revealed that the interaction was caused by more negative responses to deviant stimuli compared to standards in the 196-228 ms interval, and by more positive responses to deviant stimuli compared to standards in five time windows comprising the continuous 324-484 ms interval (all $p<0.01$). Significant main effects of ROI and Interval, as well as their interaction were also observed but not analyzed further.

---------- Figure 6 around here ----------

The ANOVA of the amplitude values for emotion deviants and standards yielded a significant interaction of Stimulus type × Interval ($F(11,363)=3.169$, $p<0.01$, $\varepsilon=0.45$, $\eta^2=0.087$). A post-hoc Tukey test revealed that the interaction was caused by more positive responses to deviant stimuli compared to standards in the 420-452 ms interval ($p<0.01$). Significant main effects of ROI and Interval, as well as their interaction were also observed but not analyzed further.

Reaction time and hit-rate

Reaction times and hit rates for the occasional changes in the fixation cross were compared between experimental blocks. Mean reaction time was 593 ms (SD=116). Analysis of variance (ANOVA) of reaction times across the 14 blocks yielded a significant effect $F(13,312)=3.78$, $p<0.025$ (Greenhouse-Geyser adjusted, $\varepsilon=0.174$), with an effect size of $\eta^2=0.14$. A post-hoc Tukey HSD test revealed that the effect was caused by the significantly longer RTs in the first block compared to the rest of the blocks ($p<0.05$), indicating rapid adjustment during the first block followed by a steady performance speed throughout the experiment.

Mean hit rate was 93.28 (SD=5.76). Analysis of variance (ANOVA) of hit rate across the 14 blocks yielded a marginally significant effect $F(13,312)=2.32$, $p<0.06$ (Greenhouse-Geyser adjusted, $\varepsilon=0.3$), with an effect size of $\eta^2=0.09$. A post-hoc Tukey HSD test revealed that the effect was caused by the significantly lower hit rate in the first block compared to blocks 8, 9, 10, 12, 13, and 14 ($p<0.05$), indicating a steady and high performance throughout the experiment following initial adjustment to the task during the first block.

Discussion

Beginning with the seminal paper by Rao and Ballard (1999), PC has become an extremely influential concept in cognitive neuroscience and currently represents one of the most compelling computational theories of perception. An experimental paradigm that was suggested early on as a suitable probe of PC in humans is the auditory MMN (Friston, 2005; Baldeweg, 2006; Stephan et al., 2006). The MMN is attractive for studies of PC, not least because the statistical structure of the stimulus sequences can be manipulated easily. This allows for straightforward tests of general predictions from PC, for example, concerning the impact of (un)predictability on ERPs. Indeed, the results from numerous auditory MMN studies are consistent with these general predictions (Wacongne et al., 2011; Schmidt et al., 2013; Phillips et al., 2015; Chennu et al., 2016; Garrido et al., 2017).

By contrast, an opportunity that has remained surprisingly unexploited is that models of PC provide formal quantities, specifically pwPEs, and predict how these should fluctuate trial-by-trial, given a particular stimulus sequence. While some sophisticated computational treatments of single-trial variations in evoked auditory and somatosensory EEG responses exist (Ostwald et al., 2012; Lieder et al., 2013b; Kolossa et al., 2015), these have either examined other potentials than MMN, were restricted to particular electrodes and time points, or used computational quantities different from pwPEs (e.g., Bayesian surprise). In the domain of visual mismatch, computational investigations have been lacking entirely so far.

To our knowledge, this study represents the first computational single-trial EEG analysis of the visual MMN. It demonstrates that visual mismatch responses reflect trial-wise pwPEs, a core quantity of PC, and thus supports the general notion that MMN can be understood as a hierarchical Bayesian inference process (Friston, 2005; Garrido et al., 2009). Specifically, we used a Bayes-optimal agent to simulate belief trajectories about probabilities of two features of human faces: color and emotion. pwPE estimates for both features showed a significant relationship to event-related potentials at the single-trial level (Fig. 3), with activations at electrodes and time windows that were comparable to classical visual MMN results (see below). Sorting single-trial ERPs according to the magnitude of the model-based pwPE estimates and selecting those with the highest and lowest pwPEs revealed the characteristic negative mismatch waveform at posterior electrodes (Fig. 4). These findings suggest that the MMN is a correlate of pwPEs as computed by a hierarchical Bayesian model. Comparing our model-based results to those obtained with traditional averaging and subtraction methods revealed that time-course and topographic distributions of the two analyses yielded highly similar results (Fig. 6).

The high hit-rate and approximately constant RT over the experiment indicates that participants complied with the task and attended the fixation cross. Hence, the pwPEs observed in our study were likely generated by an automatic mechanism that operates outside the focus of attention, in line with theories of perception as unconscious inference (Hatfield, 2002; Friston, 2005; Kiefer, 2017).

Several studies used the visual MMN to investigate neural responses to changes in color and facial emotions (see Methods). The topographical distribution and time-course of pwPEs in our current study are in line with these previous findings. However, to our knowledge, our study is the first to demonstrate that pwPEs obtained from a formal Bayesian model (HGF) are reflected by visual mismatch responses. Thus, our results represent an important advance in the interpretation of the visual MMN, elucidating the potential underlying computational processes.

Our model-based approach identified an early time window of pwPE responses in the 180-255 ms and 170-214 ms intervals for color and emotion PEs, respectively. The topographic distribution of both responses (Fig. 6B) corresponds to the topography of the known visual MMN response characterized by a posterior dominant negative potential. These intervals are also in good agreement with our current results obtained with traditional ERP analysis methods, which showed a significantly more negative response to color deviants in the 196-228 ms interval. Traditional ERP analysis did not reveal a significant mismatch response to emotion deviants in a similarly early interval, which we discuss below.

Prior studies often observed a late positive potential following the MMN peak in the deviant-minus-standard differential response dominant at the posterior scalp (Czigler et al., 2002; Zhao and Li, 2006; Czigler and Sulykos, 2010; Muller et al., 2010; Stefanics et al., 2011). Accordingly, we found significant PEs in the 320-500 ms and 405-455 ms intervals for color and emotion changes, respectively, that corresponded to positive potentials at the posterior scalp (Figs. 3 and 6). These intervals are in good agreement with the results obtained with traditional averaging and subtraction methods which revealed significant mismatch responses in the 324-484 ms and 420-452 ms intervals for color and emotion, respectively. An important result of our current study is that the 'late positive' peak also shows a significant relationship to model-based pwPE estimates. It indicates that this later potential, similar to the MMN, is also a neural correlate of PEs, despite its scalp distribution that apparently differs from that of the MMN, which suggest that different generator sources underlie the two responses. The existence of multiple significant intervals, both for color and emotion pwPEs, are in line with PC as this posits that pwPEs are minimized in sequential steps during the model update process (Friston, 2005).

A strength of our study is that the time-course and scalp topography of significant pwPE-related potentials were identified using a model-based approach that was applied to the entire time×sensor data space. This contrasts with previous studies that often restricted the statistical analysis to certain electrodes and time intervals.

We also compared our Bayesian model against a more classical alternative (change detection) to verify our computational interpretation of visual mismatch responses. This involved two GLMs incorporating either trial-wise pwPEs (from the HGF) or categorical change indices (CD model). Model comparison indicated that the pwPE model was clearly superior to the CD model in the large majority of voxels – both for a restricted mask (where both pwPE and CD models yielded significant results at the group-level) and for the entire space-time volume. Two issues are worth highlighting here. First, our Bayesian

model is generic and pwPE trajectories obtained with the HGF are unlikely to differ markedly from those generated by other Bayesian models. In fact, for any probability distribution from the exponential family, Bayesian update equations share a canonical form for precision-weighted PEs (Mathys, 2016). Second, our approach is not restricted to a particular time bin (as in Lieder et al., 2013) and does not preclude that competing models could explain different trial components differentially well. However, this potential problem of interpretability is addressed by our functionally defined mask, which is restricted to points in time-sensor space with significant mismatch responses under both models. Future extensions of the present approach could involve generative modelling of the entire waveform. While MMN waveform models do exist, these are detailed biophysical models that cannot be directly fitted to EEG data (Wacongne et al., 2012) and/or are not suited for single-trial analyses (Lieder et al., 2013a).

A limitation of our paradigm is that the necessity to control face stimuli for spatial frequency and luminance diminished details of facial expressions which are important for emotion recognition. For example, an important cue for fear, the white sclera above the pupil revealed by widely opened eyes (Darwin, 1872; Ekman and Friesen, 2003), appeared remarkably diminished after equating images for spatial frequency and luminance. This might explain why our mismatch responses to emotion changes were less robust compared to previous studies (e.g., Stefanics et al., 2012), and why our current traditional ERP analysis approach did not yield a significant mismatch response in an early time window. Although our model-based analysis revealed significant emotion pwPE responses in the early time window of 170-214 ms, the effect was mainly driven by responses to happy faces (Fig. 4D). By contrast, our model-based approach did identify significant single-trial pwPE responses to emotional faces in the early time window where visual MMN responses were observed in prior studies. This highlights advantages of using a computational modeling approach in a GLM framework at the single-subject level. First, using trial-by-trial regressors in a GLM enables us to use all trials from the experiment and hence increases the robustness of the parameter estimates whereas in traditional MMN approaches a large portion of trials are not used in the deviant vs. standard comparisons. Second, our modeling approach allowed us to include trials where *both* color and emotion changed.

Future extensions of our current work include effective connectivity analyses, such as dynamic causal modeling (DCM) that has proven useful for our understanding of the auditory MMN (e.g., Garrido et al., 2007; Moran et al., 2013, 2014; Cooray et al., 2014; Ranlund et al., 2016). Although several electrophysiological studies are consistent with propagation of pwPEs in a hierarchical network supporting PC, the interpretation is indirect and a direct embedding of computational quantities into physiological models remains to be done. Future studies may combine hierarchical Bayesian models with DCM to better characterize trial-wise computational message passing in neural circuitry mediating visual perception.

513 Table 1. Test statistics for color and emotion prediction errors.

Test statistics for color prediction errors					
Activation size (# voxels)	Cluster p-value (FEW-corrected)	Peak p-value (FWE-corrected)	Peak F-statistic	Peak Equivalent Z-statistic	Peak Latency (ms)
9885	***1.44E-10***	***2.42E-10***	40.63242	7.574789	472
		2.63E-08	32.85626	6.898473	412
		4.1E-08	32.14478	6.830418	388
		5.32E-08	31.73202	6.790405	388
3958	***4.71E-06***	***3.9E-10***	39.81532	7.50916	208
		2.11E-06	26.03102	6.192928	216
		2.6E-06	25.71183	6.156698	216
		2.5E-05	22.35084	5.753717	212
		5.9E-05	21.09702	5.592177	216
2006	***0.000426***	***4.78E-09***	35.62346	7.152807	212
9875	***1.46E-10***	***6.09E-05***	21.05077	5.586089	468
		6.31E-05	20.99963	5.579346	384
		0.000245	19.04328	5.312191	352
		0.000889	17.21467	5.044499	384
		0.002482	15.77195	4.819075	476
		0.002808	15.59909	4.791132	428
		0.003092	15.46402	4.76915	428
		0.004554	14.92295	4.679763	436
		0.010871	13.70968	4.471042	416
Test statistics for emotion prediction errors					
Activation size (# voxels)	Cluster p-value (FEW-corrected)	Peak p-value (FWE-corrected)	Peak F-statistic	Peak Equivalent Z-statistic	Peak Latency (ms)
1333	***0.001824***	***0.00334***	15.51657	4.777717	428
		0.171057	9.932535	3.729684	388
1179	***0.003041***	***0.004358***	15.14413	4.716563	188
		0.057261	11.53527	4.063691	184
		0.090418	10.87907	3.930862	180

514 Table 1. Significant activations are arranged according to size. P-values and statistics are given for
515 activation clusters and within each activation. Significant FEW-corrected p-values are in bold italics font.

516

Figure captions

Figure 1. Stimuli and paradigm. A) We used a multi-feature visual 'roving standard' paradigm to elicit PEs by rare changes of either color (red, green), or emotional expression (happy, fearful) of human faces (or both). This allowed us to study brain responses to stimuli that were physically identical but differed in whether color or emotion regularities were violated. Faces were presented in four peripheral quadrants of the screen. A detection task was presented at fixation at the center. Faces reproduced with permission of the Radboud Faces Database (www.rafd.nl). B) Schematic illustration of a stimulus sequence showing transitions between stimulus types. Note physically identical stimuli taking the role of different 'deviant' stimulus types (GH: green happy, GF: green fearful, RH: red happy, RF: red fearful faces) depending on expectations established by prior stimulus context.

Figure 2. The Hierarchical Gaussian Filter and pwPE trajectories. A) A graphical model of the Hierarchical Gaussian Filter with two levels (figure modified from Mathys et al., 2011). B) Model-based pwPE trajectories from one experimental block used as regressors in the GLM. GF: green fearful, GH: green happy, RF: red fearful, RH: red happy faces.

Figure 3. Thresholded space-time statistical parametric maps (SPMs). A) Main effects of color pwPE estimates (pooled across emotions) of the F-test (whole-scalp corrected at $p<0.05$, with a cluster-defining threshold of $p<0.001$). Crosshair is positioned at the earliest maximum of test statistics. B) Contrast estimates (arbitrary units) for the four types of stimuli (GF: green fearful, GH: green happy, RF: red fearful, RH: red happy faces) at three time points of maxima in posterior clusters. Bars indicate 90% C.I. as additional illustration for ERP effects found after whole-scalp x epoch length FWE correction. C) and D) Main effects of emotion pwPE estimates (pooled across colors) plotted similarly as for color pwPEs.

Figure 4. pwPE parameter estimates and ERP image of all single trials of 34 subjects (>283'000 single trials). Data in all subplots were smoothed with a sliding window of 3000 trials for visualization. A) Mean-centered parameter estimates of pwPEs to color input sorted from minimum (top) to maximum (bottom) values, yielded by the HGF. Data were smoothed using a vertical window of 3000 trials. B) Single-trial ERPs from occipito-temporal electrodes sorted according to their associated pwPE magnitude. Note vertical lines corresponding to ERP peaks and troughs. C) Mean ERP amplitudes over the intervals with significant correlation between pwPE and ERP. Red and purple lines show potential values averaged over the intervals 200-240 ms and 320-430 ms, respectively. Confidence intervals (S.D.) resulted from the time windows used per time point. D) ERP waveforms calculated across 10 % of trials with the lowest and highest pwPE parameter estimates. Confidence intervals (S.D.) resulted from the single trials. Note the difference between waveforms in the intervals where significant pwPE-related activity has been found with multiple regression. Red areas in head plots show scalp regions where electrodes were used for plotting the ERP waveforms. E-H) Data for emotion pwPEs plotted similarly as for color above).

Figure 5. Histograms of Δ_{LME} over the voxels within a mask defined by the conjunction of significant voxels for the pwPE and change detection models either for color or emotion changes, and over all voxels in the whole 3D space-time volume.

Figure 6. ERP waveforms, scalp voltage maps, and topographic statistical parametric maps. A) ERPs with 95% confidence interval for changes in color obtained with traditional averaging deviant-minus-standard subtraction. Red areas in channel layout plots show scalp regions where electrodes were used for plotting the ERP waveforms. B) Scalp potential plots of deviant-minus-standard difference waveform (left) at two timepoints of cluster maxima where SPM analysis yielded significant results. Statistical parametric maps (right) for model-based color pwPE estimates (pooled across emotions) of the F-test. Note high similarity of topographic distributions for the traditionally obtained mismatch responses (with negative and positive posterior scalp

distributions) and the statistical parametric map (SPM) obtained with computational model-based analyses. C-D) Data for the emotion changes, plotted similarly as for color.

References

Astikainen P, Cong F, Ristaniemi T, Hietanen JK (2013) Event-related potentials to unattended changes in facial expressions: detection of regularity violations or encoding of emotions? Front Hum Neurosci 7:557.

Astikainen P, Hietanen JK (2009) Event-related potentials to task-irrelevant changes in facial expressions. Behav Brain Funct 5:30.

Auksztulewicz R, Friston K (2015) Attentional Enhancement of Auditory Mismatch Responses: a DCM/MEG Study. Cereb Cortex 25:4273-4283.

Auksztulewicz R, Friston K (2016) Repetition suppression and its contextual determinants in predictive coding. Cortex 80:125-140.

Baldeweg T (2006) Repetition effects to sounds: evidence for predictive coding in the auditory system. Trends Cogn Sci 10:93-94.

Berg P, Scherg M (1994) A multiple source approach to the correction of eye artifacts. Electroencephalogr Clin Neurophysiol 90:229-241.

Chennu S, Noreika V, Gueorguiev D, Shtyrov Y, Bekinschtein TA, Henson R (2016) Silent Expectations: Dynamic Causal Modeling of Cortical Prediction and Attention to Sounds That Weren't. J Neurosci 36:8305-8316.

Clark A (2015) Surfing Uncertainty: Prediction, Action, and the Embodied Mind. In. Oxford: Oxford University Press.

Cooray G, Garrido MI, Hyllienmark L, Brismar T (2014) A mechanistic model of mismatch negativity in the ageing brain. Clin Neurophysiol 125:1774-1782.

Costa-Faidella J, Baldeweg T, Grimm S, Escera C (2011a) Interactions between "what" and "when" in the auditory system: temporal predictability enhances repetition suppression. J Neurosci 31:18590-18597.

Costa-Faidella J, Grimm S, Slabu L, Diaz-Santaella F, Escera C (2011b) Multiple time scales of adaptation in the auditory system as revealed by human evoked potentials. Psychophysiol 48:774-783.

Csukly G, Stefanics G, Komlosi S, Czigler I, Czobor P (2013) Emotion-related visual mismatch responses in schizophrenia: impairments and correlations with emotion recognition. PLoS One 8:e75444.

Czigler I, Balazs L, Winkler I (2002) Memory-based detection of task-irrelevant visual changes. Psychophysiology 39:869-873.

Czigler I, Sulykos I (2010) Visual mismatch negativity to irrelevant changes is sensitive to task-relevant changes. Neuropsychol 48:1277-1282.

Darwin C (1872) The expression of emotions in man and animals. London: John Murray.

Ekman P, Friesen WV (2003) Unmasking the Face; a Guide to Recognizing Emotions from Facial Clues. Cambridge, MA.: Malor Books.

Flandin G, Friston KJ (2016) Analysis of family-wise error rates in statistical parametric mapping using random field theory. arXiv:160608199v1.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815-836.

Garrido MI, Friston KJ, Kiebel SJ, Stephan KE, Baldeweg T, Kilner JM (2008) The functional anatomy of the MMN: a DCM study of the roving paradigm. Neuroimage 42:936-944.

Garrido MI, Kilner JM, Kiebel SJ, Friston KJ (2009) Dynamic causal modeling of the response to frequency deviants. J Neurophysiol 101:2620-2631.

Garrido MI, Kilner JM, Kiebel SJ, Stephan KE, Friston KJ (2007) Dynamic causal modelling of evoked potentials: a reproducibility study. Neuroimage 36:571-580.

Garrido MI, Rowe EG, Halász V, Mattingley JB (2017) Bayesian Mapping Reveals That Attention Boosts Neural Responses to Predicted and Unpredicted Stimuli. Cereb Cortex:1-12.

Garrido MI, Sahani M, Dolan RJ (2013) Outlier responses reflect sensitivity to statistical structure in the human brain. PLoS Comput Biol 9:e1002999.

Haenschel C, Vernon DJ, Dwivedi P, Gruzelier JH, Baldeweg T (2005) Event-related brain potential correlates of human auditory sensory memory-trace formation. J Neurosci 25:10494-10501.

Hatfield G (2002) Perception as Unconscious Inference. In: Perception and the Physical World: Psychological and Philosophical Issue in Perception (Heyer D, Mausfeld R, eds), pp 115-143: John Wiley & Sons.

Hauser TU, Iannaccone R, Ball J, Mathys C, Brandeis D, Walitza S, Brem S (2014) Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. JAMA Psychiatry 71:1165-1173.

Henson RN, Mouchlianitis E, Matthews WJ, Kouider S (2008) Electrophysiological correlates of masked face priming. Neuroimage 40:884-895.

Hohwy J (2013) The predictive mind. Oxford: Oxford University Press.

Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HE, Stephan KE (2013) Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron 80:519-530.

Jepma M, Murphy PR, Nassar MR, Rangel-Gomez M, Meeter M, Nieuwenhuis S (2016) Catecholaminergic Regulation of Learning Rate in a Dynamic Environment. PLoS Comput Biol 12:e1005171.

Kass RE, Raftery AE (1995) Bayes Factors. JASA 90:773-795.

Kiebel SJ, Friston KJ (2004) Statistical parametric mapping for event-related potentials: I. Generic considerations. Neuroimage 22:492-502.

Kiefer A (2017) Literal Perceptual Inference. In: Philosophy and Predictive Processing: 17 (Metzinger T, Wiese W, eds). Frankfurt am Main: MIND Group.

Kimura M, Katayama J, Murohashi H (2006) Probability-independent and -dependent ERPs reflecting visual change detection. Psychophysiology 43:180-189.

Kimura M, Kondo H, Ohira H, Schröger E (2012) Unintentional temporal context-based prediction of emotional faces: an electrophysiological study. Cereb Cortex 22:1774-1785.

Kolossa A, Kopp B, Fingscheidt T (2015) A computational analysis of the neural bases of Bayesian inference. Neuroimage 106:222-237.

Komatsu M, Takaura K, Fujii N (2015) Mismatch negativity in common marmosets: Whole-cortical recordings with multi-channel electrocorticograms. Sci Rep 5:15006.

Kreegipuu K, Kuldkepp N, Sibolt O, Toom M, Allik J, Näätänen R (2013) vMMN for schematic faces: automatic detection of change in emotional expression. Front Hum Neurosci 7:714.

Kremláček J, Kreegipuu K, Tales A, Astikainen P, Põldver N, Näätänen R, Stefanics G (2016) Visual mismatch negativity (vMMN): A review and meta-analysis of studies in psychiatric and neurological disorders. Cortex 80:76-112.

Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A (2010) Presentation and validation of the Radboud Faces Database. Cognition & Emotion 24:1377-1388.

Lawson RP, Mathys C, Rees G (2017) Adults with autism overestimate the volatility of the sensory environment. Nat Neurosci 20:1293-1299.

Lieder F, Daunizeau J, Garrido MI, Friston KJ, Stephan KE (2013b) Modelling trial-by-trial changes in the mismatch negativity. PLoS Comput Biol 9:e1002911.

Lieder F, Stephan KE, Daunizeau J, Garrido MI, Friston KJ (2013a) A neurocomputational model of the mismatch negativity. PLoS Comput Biol 9:e1003288.

Litvak V, Mattout J, Kiebel S, Phillips C, Henson R, Kilner J, Barnes G, Oostenveld R, Daunizeau J, Flandin G, Penny W, Friston K (2011) EEG and MEG data analysis in SPM8. Comput Intell Neurosci 2011:852961.

Mathys C (2016) How Could We Get Nosology from Computation? In: Computational Psychiatry: New Perspectives on Mental Illness (Redish AD, Gordon JA, eds). Strüngmann Forum Reports, vol. 20. Cambridge, MA: MIT Press.

Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A bayesian foundation for individual learning under uncertainty. Front Hum Neurosci 5:39.

Mathys C, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, Stephan KE (2014) Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci 8:825.

Mo L, Xu G, Kay P, Tan LH (2011) Electrophysiological evidence for the left-lateralized effect of language on preattentive categorical perception of color. Proc Natl Acad Sci U S A 108:14026-14030.

Moran RJ, Campo P, Symmonds M, Stephan KE, Dolan RJ, Friston KJ (2013) Free energy, precision and learning: the role of cholinergic neuromodulation. J Neurosci 33:8227-8236.

Moran RJ, Symmonds M, Dolan RJ, Friston KJ (2014) The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. PLoS Comput Biol 10:e1003422.

Müller D, Winkler I, Roeber U, Schaffer S, Czigler I, Schröger E (2010) Visual Object Representations Can Be Formed outside the Focus of Voluntary Attention: Evidence from Event-related Brain Potentials. J Cogn Neurosci 22:1179-1188.

Näätänen R, Astikainen P, Ruusuvirta T, Huotilainen M (2010) Automatic auditory intelligence: an expression of the sensory-cognitive core of cognitive processes. Brain Res Rev 64:123-136.

Näätänen R, Kujala T, Escera C, Baldeweg T, Kreegipuu K, Carlson S, Ponton C (2012) The mismatch negativity (MMN)-a unique window to disturbed central auditory processing in ageing and different clinical conditions. Clin Neurophysiol 123:424-458.

Näätänen R, Paavilainen P, Alho K, Reinikainen K, Sams M (1989) Do Event-Related Potentials Reveal the Mechanism of the Auditory Sensory Memory in the Human Brain. Neurosci Lett 98:217-221.

Näätänen R, Paavilainen P, Tiitinen H, Jiang D, Alho K (1993) Attention and mismatch negativity. Psychophysiology 30:436-450.

Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I (2001) 'Primitive intelligence' in the auditory cortex. Trends Neurosci 24:283-288.

672 Ostwald D, Spitzer B, Guggenmos M, Schmidt TT, Kiebel SJ, Blankenburg F (2012) Evidence for neural encoding of Bayesian
673 surprise in human somatosensation. Neuroimage 62:177-188.
674 Paavilainen P (2013) The mismatch-negativity (MMN) component of the auditory event-related potential to violations of
675 abstract regularities: a review. Int J Psychophysiol 88:109-123.
676 Paavilainen P, Arajarvi P, Takegata R (2007) Preattentive detection of nonsalient contingencies between auditory features.
677 Neuroreport 18:159-163.
678 Phillips HN, Blenkmann A, Hughes LE, Bekinschtein TA, Rowe JB (2015) Hierarchical Organization of Frontotemporal Networks
679 for the Prediction of Stimuli across Multiple Dimensions. J Neurosci 35:9255-9264.
680 Powers AR, Mathys C, Corlett PR (2017) Pavlovian conditioning-induced hallucinations result from overweighting of perceptual
681 priors. Science 357:596-600.
682 Preuschoff K, Bossaerts P (2007) Adding prediction risk to the theory of reward learning. Ann N Y Acad Sci 1104:135-146.
683 Ranlund S, Adams RA, Diez A, Constante M, Dutt A, Hall MH, Maestro Carbayo A, McDonald C, Petrella S, Schulze K, Shaikh M,
684 Walshe M, Friston K, Pinotsis D, Bramon E (2016) Impaired prefrontal synaptic gain in people with psychosis and their
685 relatives during the mismatch negativity. Hum Brain Mapp 37:351-365.
686 Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-
687 field effects. Nat Neurosci 2:79-87.
688 Rescorla RA, Wagner AR (1972) A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and
689 Nonreinforcement. In: Classical Conditioning II: Current Research and Theory (Black AH, Prokasy WF, eds), pp 64–99.
690 New York: Appleton-Century-Crofts.
691 Schmidt A, Diaconescu AO, Kometer M, Friston KJ, Stephan KE, Vollenweider FX (2013) Modeling ketamine effects on synaptic
692 plasticity during the mismatch negativity. Cereb Cortex 23:2394-2406.
693 Schröger E (1998) Measurement and interpretation of the mismatch negativity. Behav Res Meth Ins C 30:131-145.
694 Schwartenbeck P, FitzGerald TH, Mathys C, Dolan R, Friston K (2015) The Dopaminergic Midbrain Encodes the Expected
695 Certainty about Desired Outcomes. Cereb Cortex 25:3434-3445.
696 Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6:461-464.
697 Stefanics G, Astikainen P, Czigler I (2015) Visual mismatch negativity (vMMN): a prediction error signal in the visual modality.
698 Front Hum Neurosci 8.
699 Stefanics G, Csukly G, Komlosi S, Czobor P, Czigler I (2012) Processing of unattended facial emotions: a visual mismatch
700 negativity study. Neuroimage 59:3042-3049.
701 Stefanics G, Czigler I (2012) Automatic prediction error responses to hands with unexpected laterality: an electrophysiological
702 study. Neuroimage 63:253-261.
703 Stefanics G, Haden GP, Sziller I, Balazs L, Beke A, Winkler I (2009) Newborn infants process pitch intervals. Clin Neurophysiol
704 120:304-308.
705 Stefanics G, Kimura M, Czigler I (2011) Visual mismatch negativity reveals automatic detection of sequential regularity violation.
706 Front Hum Neurosci 5:46.
707 Stefanics G, Kremlacek J, Czigler I (2014) Visual mismatch negativity: a predictive coding view. Front Hum Neurosci 8:666.
708 Stephan KE, Baldeweg T, Friston KJ (2006) Synaptic plasticity and dysconnection in schizophrenia. Biol Psychiatry 59:929-939.
709 Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage
710 46:1004-1017.
711 Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007) Comparing hemodynamic models with DCM. Neuroimage
712 38:387-401.
713 Takaura K, Fujii N (2016) Facilitative effect of repetitive presentation of one stimulus on cortical responses to other stimuli in
714 macaque monkeys--a possible neural mechanism for mismatch negativity. Eur J Neurosci 43:516-528.
715 Thierry G, Athanasopoulos P, Wiggett A, Dering B, Kuipers JR (2009) Unconscious effects of language-specific terminology on
716 preattentive color perception. P Natl Acad Sci USA 106:4567-4570.
717 Vossel S, Mathys C, Stephan KE, Friston KJ (2015) Cortical Coupling Reflects Bayesian Belief Updating in the Deployment of
718 Spatial Attention. J Neurosci 35:11532-11542.
719 Wacongne C, Changeux JP, Dehaene S (2012) A neuronal model of predictive coding accounting for the mismatch negativity. J
720 Neurosci 32:3665-3678.
721 Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S (2011) Evidence for a hierarchy of predictions
722 and prediction errors in human cortex. Proc Natl Acad Sci U S A 108:20754-20759.
723 Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW (2010) Controlling low-level image properties: the SHINE
724 toolbox. Behav Res Methods 42:671-684.
725 Winkler I (2007) Interpreting the Mismatch Negativity. J Psychophysiol 21:147-163.

726 Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC (1996) A unified statistical approach for determining significant
727 signals in images of cerebral activation. Hum Brain Mapp 4:58-73.
728 Zhao L, Li J (2006) Visual mismatch negativity elicited by facial expressions under non-attentional condition. Neurosci Lett
729 410:126-131.

730